Dan Hendrycks

UC Berkeley
hendrycks@berkeley.edu
www.danhendrycks.com

Current Position

PhD student in computer science *2019–Present*
UC Berkeley

Education

UC Berkeley, PhD *Fall 2018 – present*
Computer Science

University of Chicago, B.S. with Honors *2018*
Computer Science

Marshfield High School in Missouri, Valedictorian *2014*

Honors & Awards

NSF GRFP Fellowship *2019 -*

Open Philanthropy Project AI Fellow *2019 -*

Selected Invited Talks

DeepMind FHI AI Safety Seminar *September 20, 2021*
Unsolved Safety Problems

Vision for All Seasons Workshop *June 25, 2021*
Talk on Robustness and Tail Risks (CVPR 2021 Workshop)

Center for Democracy & Technology *May 27, 2021*
The Capabilities and Limitations of Automated Content Analysis (Panel)

Radboud University Medical Center, Netherlands *March 24, 2021*
OOD Detection

OpenAI *March 3, 2021*
Safety vs. Capabilities

VITA Group (UT Austin) *February 21, 2021*
Robustness

Workshop on Dataset Curation and Security *December 11, 2021*
Talk on Robustness (NeurIPS 2020 Workshop)

Data Fest `https://fest.ai` *October 10, 2020*
Tutorial on OOD Detection

Hazy Research (Stanford) *September 22, 2020*
Data Augmentation and Massive Multitask Learning

Perth ML Reading Group *August 27, 2020*
Value Learning and Machine Ethics

UberAI *July 31, 2020*
Robustness in Vision and NLP

DeepMind *April 20, 2020*
OOD Detection

BBC World News Live Interview *June 26, 2019*
Adversarially Filtered Images

Publications

[21] **Hendrycks, D.**, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization". In: *ICCV* (2021). URL: https://arxiv.org/abs/2006.16241.

[20] **Hendrycks, D.**, K. Zhao, S. Basart, J. Steinhardt, and D. Song. "Natural Adversarial Examples". In: *CVPR* (2021). URL: http://arxiv.org/abs/1907.07174.

[19] **Hendrycks, D.**, S. Basart, S. Kadavath, M. Mazeika, A. Arora, E. Guo, C. Burns, S. Puranik, H. He, D. Song, and J. Steinhardt. "Measuring Coding Challenge Competence With APPS". In: *arXiv preprint arXiv:2105.09938* (2021). URL: https://arxiv.org/abs/2105.09938.

[18] **Hendrycks, D.**, C. Burns, A. Chen, and S. Ball. "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review". In: *arXiv preprint arXiv:2103.06268* (2021). URL: https://arxiv.org/abs/2103.06268.

[17] **Hendrycks, D.**, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. "Measuring Mathematical Problem Solving With the MATH Dataset". In: *arXiv preprint arXiv:2103.03874* (2021). URL: https://arxiv.org/abs/2103.03874.

[16] **Hendrycks, D.**, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. "Measuring Massive Multitask Language Understanding". In: *ICLR* (2021). URL: https://arxiv.org/abs/2009.03300.

[15] **Hendrycks, D.**, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. "Aligning AI With Shared Human Values". In: *ICLR* (2021). URL: https://arxiv.org/abs/2008.02275.

[14] **Hendrycks, D.**, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. "Pretrained Transformers Improve Out-of-Distribution Robustness". In: *ACL* (2020). URL: https://arxiv.org/abs/2004.06100.

[13] **Hendrycks, D.**, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty". In: *ICLR* (2020). URL: https://arxiv.org/abs/1912.02781.

[12] **Hendrycks, D.**, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song. "Scaling Out-of-Distribution Detection for Real-World Settings". In: *arXiv preprint arXiv:1911.11132* (2019). URL: https://arxiv.org/abs/1911.11132.

[11] **Hendrycks, D.**, M. Mazeika, S. Kadavath, and D. Song. "Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty". In: *NeurIPS* (2019). URL: http://arxiv.org/abs/1906.12340.

[10] Kang, D., Y. Sun, **D. Hendrycks**, T. Brown, and J. Steinhardt. "Testing Robustness Against Unforeseen Adversaries". In: *arXiv preprint arXiv:1908.08016* (2019). URL: http://arxiv.org/abs/1908.08016.

[9] Gilmer, J. and **D. Hendrycks**. "A Discussion of 'Adversarial Examples Are Not Bugs, They Are Features': Adversarial Example Researchers Need to Expand What is Meant by 'Robustness'". In: *Distill* (2019). URL: https://doi.org/10.23915/distill.00019.1.

[8] **Hendrycks, D.**, K. Lee, and M. Mazeika. "Using Pre-Training Can Improve Model Robustness and Uncertainty". In: *ICML* (2019). URL: http://arxiv.org/abs/1901.09960.

[7] **Hendrycks, D.**, M. Mazeika, and T. G. Dietterich. "Deep Anomaly Detection with Outlier Exposure". In: *ICLR* (2019). URL: http://arxiv.org/abs/1812.04606.

[6] **Hendrycks, D.** and T. G. Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *ICLR* (2019). URL: http://arxiv.org/abs/1903.12261.

[5] **Hendrycks, D.**, M. Mazeika, D. Wilson, and K. Gimpel. "Using Trusted Data to Train Deep Networks on Labels Corrupted by Severe Noise". In: *NeurIPS* (2018). URL: http://arxiv.org/abs/1802.05300.

[4] Liu, S., R. Garrepalli, T. G. Dietterich, A. Fern, and **D. Hendrycks**. "Open Category Detection with PAC Guarantees". In: *ICML* (2018). URL: http://arxiv.org/abs/1808.00529.

[3] **Hendrycks, D.** and K. Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *ICLR* abs/1610.02136 (2016). arXiv: 1610.02136. URL: http://arxiv.org/abs/1610.02136.

[2] **Hendrycks, D.** and K. Gimpel. "Early Methods for Detecting Adversarial Images". In: *ICLR Workshop* (2016). URL: http://arxiv.org/abs/1608.00530.

[1] **Hendrycks, D.** and K. Gimpel. "Gaussian Error Linear Units (GELUs)". In: *Technical Report* (2016). URL: http://arxiv.org/abs/1606.08415.

Experience

DeepMind Research Intern *May. 2019 - Aug. 2019*

GiveWell Summer Fellow *Aug. 2015*

Service

- I co-organized the Workshop on Robustness and Uncertainty Estimation in Deep Learning at ICML 2019, 2020, 2021
- I co-organized the Adversarial Machine Learning workshop at ICML 2021
- I co-organzied the VISDA domain adaptation competition at NeurIPS 2021
- I have reviewed for CVPR (2019, 2020), ICLR (2019, 2020), NeurIPS (2017, 2020), ICML (2018, 2019), ICCV (2019), ECCV (2018), IJCV, TPAMI, and JMLR.